Plithogenic Tensor Neutrosophic Logic: A Formal Foundation for Declared-Loss Evaluation in Large Language Models

Maikel Yelandi Leyva-Vázquez[1] and Florentin Smarandache[2]

[1]Facultad de Ciencias Matemáticas y Físicas, Universidad de Guayaquil, Ecuador; Universidad Bolivariana del Ecuador, Durán, Ecuador; Universidad Bernardo O'Higgins, Santiago, Chile. mleyvaz@gmail.com. ORCID: 0000-0001-7911-5879.
[2]Mathematics, Physics, and Natural Sciences Division, University of New Mexico, Gallup, NM 87301, USA. smarand@unm.edu

April 2026

Abstract

Mason (2026) replicates the neutrosophic Truth/Indeterminacy/Falsity (T/I/F) evaluation protocol of Leyva-Vázquez and Smarandache (2025) across five large-language-model vendors and reports cross-vendor hyper-truth in 84% of unconstrained evaluations. He further identifies an *Absorption Problem*: several models map paradox, ignorance, and contingency to the identical scalar (T=0, $I\approx$1, F=0), collapsing precisely the distinctions neutrosophic logic was introduced to preserve. His proposed remedy is a tensor extension — the S4 protocol — in which each evaluation is augmented with a list of *declared losses* of the form (what, why, severity). The losses discriminate phenomena that the scalars equate (Jaccard similarity < 0.10 for paradox versus ignorance across all five models). We contribute a formal foundation for this empirical advance. We show that S4 outputs are already, in a precise mathematical sense, single-valued plithogenic neutrosophic structures (Smarandache, 2018), and that the Absorption Problem dissolves under the standard plithogenic decomposition by sub-attributes. We give a unified plithogenic distance metric that combines scalar Manhattan distance, attribute-set Jaccard or embedding distance, and a contradiction-function term, recovering the discriminations Mason reports. We compare plithogeny with Dempster–Shafer evidence theory, Type-2 fuzzy sets, modal logic, rough sets, case-based reasoning, and Atanassov intuitionistic fuzzy sets, and argue that plithogeny uniquely satisfies the conjunction of constraints required by S4. Finally, we connect the framework to our prior work on dynamic epistemic neutrosophic logic (LED), neutrosophic analysis of competing hypotheses (N-ACH), neutrosophic explainable AI (NeutroXAI), generative functional decomposition (GFD-N), and neutrosophic dynamic Bayesian graphs (NDBG).

1 Introduction

Neutrosophic logic, as introduced by smarandache2005, smarandache2010, treats Truth (T), Indeterminacy (I), and Falsity (F) as three independent components of an epistemic state, freed from the probabilistic constraint $T + I + F = 1$. In leyva2025 we showed that this freedom matters in practice: when large language models are prompted to produce (T, I, F) triples for paradoxical, ignorance-based, vague, ethical-conflict, and future contingent statements, 35% of evaluations exceed the probabilistic sum, exhibiting hyper-truth $T+I+F > 1$. Hyper-truth is not noise; it is the model expressing internal conflict that probabilistic output formats systematically suppress.

mason2026 extends this finding in two complementary directions. First, he replicates the experiment across five model families from five vendors — Anthropic, Meta, DeepSeek, Alibaba,

Mistral — with five repetitions per cell, finding hyper-truth in 84% of unconstrained evaluations and confirming that the phenomenon is cross-vendor rather than an OpenAI-specific artifact. Second, and more interestingly for the present paper, he identifies a genuine representational ceiling of the scalar (T, I, F) output. Several models, most prominently Mistral and the original GPT-4o, adopt what he calls an *Absorption* position, mapping paradox, ignorance, and contingency to the same scalar ($T{=}0, I{\approx}1, F{=}0$). The distinctions that neutrosophic logic was designed to preserve collapse precisely on the cases where they matter most.

Mason's response is to extend the output format. His S4 protocol asks the model not only for a (T, I, F) triple but also for a list of *declared losses*, each of the form

$$\ell_j = (\mathsf{what}_j, \mathsf{why}_j, \mathsf{severity}_j), \qquad \mathsf{severity}_j \in [0, 1].$$

Empirically the losses discriminate where the scalars do not: for paradox versus ignorance, the word-level Jaccard similarity between loss vocabularies is below 0.10 for all five models, regardless of whether the scalar Manhattan distance is zero. Mason summarizes the finding crisply: *the model has the distinction; the scalar cannot express it; the tensor can.*

This paper provides what Mason's empirical work invites: a formal foundation for the S4 tensor. We argue that S4 is not a novel representational structure but an instance of a pre-existing one — the plithogenic neutrosophic structure of smarandache2018. Under the natural mapping $\mathsf{what} \mapsto$ attribute value, $\mathsf{severity} \mapsto$ membership degree, $\mathsf{why} \mapsto$ data for the contradiction function, the S4 output of Mason becomes a single-valued plithogenic neutrosophic set. The Absorption Problem dissolves because two evaluations with identical scalar projections may correspond to plithogenic structures with disjoint attribute spectra. Mason's Jaccard finding is, on this reading, an empirical estimator of one minus the plithogenic contradiction function.

The contribution is fivefold:

1. We give a precise embedding of S4 into the plithogenic framework (§3, Theorem 3.1).
2. We resolve the Absorption Problem formally (§4, Proposition 4.2).
3. We propose a unified plithogenic distance metric and show that Mason's reported similarities instantiate it (Definition 4.3).
4. We compare plithogeny against six alternative frameworks and argue that only plithogeny satisfies the conjunction of constraints required by S4 (§5).
5. We connect the framework to our prior work on LED, N-ACH, NeutroXAI, GFD-N, and NDBG, showing that each is an instance or extension of the same plithogenic backbone (§6).

We do not contest Mason's empirical findings; we extend them by embedding his tensor in a richer mathematical structure that has been in the literature since 2018. The closing section invites empirical collaboration on a strengthened S4 protocol, which we denote S4-N, that records neutrosophic per-attribute severities and exposes the contradiction function explicitly.

2 Background: Plithogenic Neutrosophic Logic

We summarize the formal apparatus we will need. The exposition follows smarandache2018, with notation lightly adapted to the LLM evaluation setting.

Definition 2.1 (Plithogenic structure)**.** *A* plithogenic structure *is a 5-tuple* $\mathcal{P} = (P, v, V, d, c)$ *where*

- P *is a set of* plithogenic elements *(in our setting, statements under evaluation by an LLM),*
- v *is the* dominant attribute*,*

- V *is the* spectrum of attribute values*, finite or continuous,*
- $d : P \times V \to [0,1]^n$ *is the* degree-of-appurtenance function *(with $n = 1$ fuzzy, $n = 2$ intuitionistic, $n = 3$ neutrosophic),*
- $c : V \times V \to [0,1]$ *is the* contradiction function*, satisfying $c(v,v) = 0$ and $c(v_i, v_j) = c(v_j, v_i)$.*

Definition 2.2 (Single-valued plithogenic neutrosophic set)**.** *A* single-valued plithogenic neutrosophic set *is a plithogenic structure with $n = 3$ and $d(p,v) = (T_v(p), I_v(p), F_v(p)) \in [0,1]^3$ for each $p \in P$ and $v \in V$.*

The contradiction function c is the formal device that distinguishes plithogeny from earlier multi-valued logics. Two attributes with $c(v_i, v_j) = 0$ are taken as semantically synonymous; with $c = 1$ they are taken as maximally opposed. Operationally, c may be estimated by lexical overlap (Jaccard or token cosine), by sentence-embedding distance under a contemporary encoder, or by domain-expert elicitation. We return to operationalization in §3.

The plithogenic operators on a single-valued neutrosophic structure take the schematic form

$$(d_1 \wedge_{\mathcal{P}} d_2)(p,v) = w_c\, T_v(p_1) \otimes T_v(p_2) \oplus (1 - w_c)\, T_v(p_1) \otimes_F T_v(p_2), \tag{1}$$

$$(d_1 \vee_{\mathcal{P}} d_2)(p,v) = w_c\, T_v(p_1) \oplus T_v(p_2) \otimes (1 - w_c)\, T_v(p_1) \oplus_F T_v(p_2), \tag{2}$$

where w_c is a weighting derived from c between the relevant attribute values, $\otimes$ is a t-norm, and $\oplus$ is its dual t-conorm; an analogous formula holds component-wise for I and F. The point we wish to emphasize is not the algebraic detail but the fact that plithogeny gives the operators access to the contradiction structure c. Operations between attributes are not flat; they are weighted by semantic relatedness.

Remark 2.3. *Definition 2.1 subsumes classical fuzzy sets ($n = 1$, $|V| = 1$, $c \equiv 0$), intuitionistic fuzzy sets ($n = 2$, $|V| = 1$), and the standard neutrosophic set ($n = 3$, $|V| = 1$). It generalizes by allowing V to be larger than one and c to be non-trivial.*

3 Mason's S4 as Plithogenic Evaluation

We now embed the S4 protocol of mason2026 into the plithogenic framework. Recall that for each statement s, S4 elicits a tuple

$$\mathrm{S4}(s) = \big(T(s), I(s), F(s), \{\ell_1, \ldots, \ell_k\}\big), \quad \ell_j = (\mathsf{what}_j, \mathsf{why}_j, \mathsf{severity}_j). \tag{3}$$

Theorem 3.1 (Plithogenic embedding of S4)**.** *For each statement s, the S4 output corresponds to the single-valued plithogenic neutrosophic structure $\mathcal{P}_s = (P_s, v, V_s, d_s, c_s)$ defined by*

$$\begin{aligned} P_s &= \{s\}, \\ v &= \text{“epistemic limitation”}, \\ V_s &= \{\mathsf{what}_1, \ldots, \mathsf{what}_k\}, \\ d_s(s, \mathsf{what}_j) &= \big(T(s), I(s) \cdot \mathsf{severity}_j, F(s)\big) \in [0,1]^3, \\ c_s(\mathsf{what}_i, \mathsf{what}_j) &\in [0,1] \text{ obtained from } \mathsf{why}_i, \mathsf{why}_j. \end{aligned}$$

The induced plithogenic operators agree with the natural composition rules on S4 outputs.

Proof sketch. The mapping is well-defined because severity is bounded in $[0,1]$ and $T, F \in [0,1]$. Closure under $\wedge_{\mathcal{P}}$ and $\vee_{\mathcal{P}}$ follows from the closure of t-norms and t-conorms on the unit cube. The contradiction function c_s is admissible whenever the operationalization (Jaccard, embeddings, or expert elicitation) returns values in $[0,1]$ with $c(v,v) = 0$ and symmetric. □

Remark 3.2 (Operationalization of c_s)**.** *Mason reports word-level Jaccard similarity on* what *fields. We propose two complementary estimators, both supported by the data he releases. Let W_i, W_j be the bags of tokens of* what_i *and* what_j*, and let $\phi(\mathsf{why}_i)$ be the sentence embedding of* why_i *under a fixed encoder (e.g. all-MiniLM-L6-v2). Define*

$$c_s^{lex}(v_i, v_j) = 1 - J(W_i, W_j), \qquad J(W_i, W_j) = \frac{|W_i \cap W_j|}{|W_i \cup W_j|}, \tag{4}$$

$$c_s^{sem}(v_i, v_j) = 1 - \frac{\phi(\mathsf{why}_i)^\top \phi(\mathsf{why}_j)}{\|\phi(\mathsf{why}_i)\| \|\phi(\mathsf{why}_j)\|}. \tag{5}$$

Both are bounded in $[0, 1]$ and symmetric, both vanish on $v_i = v_j$, and either may be substituted into Definition 2.1. The lexical estimator is what Mason effectively reports; the semantic estimator is more discriminating when the loss vocabularies are synonymic but lexically disjoint.

Worked examples. We re-express two of Mason's most striking observations in plithogenic notation.

Mistral on the liar paradox. Mason reports the modal S4 output $(T, I, F) = (0.0, 1.0, 0.0)$ with three losses: "self-referential paradox resolution" (severity 1.0), "formal-system dependency" (0.8), and "contextual grounding of *this sentence*" (1.0). The plithogenic structure is

$$\begin{aligned} P_{\text{liar}} &= \{\text{“This sentence is false.”}\}, \\ V_{\text{liar}} &= \{\text{self-ref}, \text{form-sys}, \text{ctx-ground}\}, \\ d_{\text{liar}}(\cdot, v) &= (0.0,\ 1.0 \cdot s_v,\ 0.0), \quad s_v \in \{1.0, 0.8, 1.0\}. \end{aligned}$$

Mistral on epistemic ignorance ("the number of stars in the universe is even"). Same scalar $(0.0, 1.0, 0.0)$ but disjoint losses: "empirical unknowability" (1.0), "definition of universe in cosmological context" (0.9), "mathematical ambiguity of *even* for infinite quantities" (0.8). The plithogenic structure is $V_{\text{ign}} = \{\text{emp-unknown}, \text{def-univ}, \text{math-amb-inf}\}$ with $V_{\text{liar}} \cap V_{\text{ign}} = \emptyset$.

The two structures share the scalar projection $\pi(\mathcal{P}) = (T, I, F)$ but differ at the attribute level. They are formally non-isomorphic plithogenic objects.

Corollary 3.3. *Theorem 3.1 embeds S4 as the special case in which* severity *is scalar in $[0, 1]$. The full plithogenic framework permits per-attribute neutrosophic severity $(T_v, I_v, F_v) \in [0, 1]^3$. We refer to the protocol that elicits this richer output as* S4-N*, building on our earlier work on NeutroXAI leyvaNeutroXAI where component-wise neutrosophic attribution was introduced.*

4 Resolving the Absorption Problem

The Absorption Problem in Mason's terminology is the empirical observation that several models map distinct epistemic states — paradox, ignorance, contingency — to the same scalar $(T, I, F) \approx (0, 1, 0)$. The scalar projection collapses the distinction.

Remark 4.1 (The Absorption Problem is an LLM behaviour artefact, not a NL limitation)**.** *A preliminary clarification is warranted. The vector $(T{=}0, I{=}1, F{=}0)$ is a perfectly valid neutrosophic triple; neutrosophic logic imposes no constraint that prevents it. More importantly, NL already provides the scalar resources to distinguish paradox from ignorance without any tensor extension: a logically paradoxical statement warrants positive T and positive F alongside maximal I, yielding, e.g.,* (0.5, 1.0, 0.5) *with $T + I + F = 2.0$, whereas genuine ignorance warrants* (0, 1, 0)*. These are distinct triples, and NL expresses both naturally. Mason's own data confirm this: Claude Sonnet produces exactly (T=0.5, I=1.0, F=0.5) for the liar paradox across all five repetitions (his* Position 1: Saturation*), which is precisely the scalar distinction that NL affords. The Absorption Problem is therefore* not *a theoretical limitation of neutrosophic logic; it is an*

empirical observation about how certain LLMs implement NL under specific prompt conditions. Those models fail to exercise the full expressive range the formalism already provides. The plithogenic tensor framework does not add expressiveness absent from scalar NL; it provides evaluation structure that guides LLMs toward using that range.

Proposition 4.2 (Distinction preserved under plithogenic embedding)**.** *Let $\pi : \mathcal{P} \mapsto (T, I, F)_{\text{global}}$ be the scalar projection, where the global triple is obtained by an attribute-aggregation rule (e.g., severity-weighted mean of $d(p, v)$). Suppose $p_1, p_2 \in P$ satisfy $\pi(\mathcal{P}_{p_1}) = \pi(\mathcal{P}_{p_2})$ but $V_{p_1} \cap V_{p_2} = \emptyset$. Then $\mathcal{P}_{p_1} \ncong \mathcal{P}_{p_2}$ as plithogenic structures, and there exists a metric $d_{\mathcal{P}}$ on plithogenic structures with $d_{\mathcal{P}}(\mathcal{P}_{p_1}, \mathcal{P}_{p_2}) > 0$.*

Proof. Plithogenic isomorphism requires a bijection on V preserving d and c. With $V_{p_1} \cap V_{p_2} = \emptyset$, no such bijection respects the empirical evidence (loss labels and contradiction estimates are stimulus-specific in Mason's data, with cross-stimulus Jaccard exactly 0.000, see his Section 4.6.2). The metric $d_{\mathcal{P}}$ of Definition 4.3 below witnesses the inequality. □

Definition 4.3 (Unified plithogenic distance)**.** *For two plithogenic structures $\mathcal{P}_1, \mathcal{P}_2$ over a common universe of statements, define*

$$d_{\mathcal{P}}(\mathcal{P}_1, \mathcal{P}_2) \;=\; \alpha\, d_{\text{esc}}(\mathcal{P}_1, \mathcal{P}_2) + \beta\,(1 - J(V_1, V_2)) + \gamma\,\bar{c}\,(V_1 \cup V_2), \tag{6}$$

where $d_{\text{esc}}(\mathcal{P}_1, \mathcal{P}_2) = |T_1 - T_2| + |I_1 - I_2| + |F_1 - F_2|$ is the scalar Manhattan distance on the projected triples, $J(V_1, V_2)$ is the Jaccard similarity of the attribute spectra, and $\bar{c}\,(V_1 \cup V_2)$ is the average contradiction across cross-set pairs $\{(v_i, v_j) : v_i \in V_1, v_j \in V_2\}$. The weights satisfy $\alpha, \beta, \gamma \geq 0$ and $\alpha + \beta + \gamma = 1$, and may be calibrated empirically.

Consistency with Mason's data. Mason's reported scalar Manhattan distances of 0.034 (Claude), 0.040 (Llama), and 0.000 (Mistral) for paradox versus ignorance yield d_{esc} near zero. His reported Jaccard similarities of 0.097 (Claude), 0.056 (Llama), and 0.066 (Mistral) yield attribute-distance contributions $1 - J \in [0.90, 0.94]$. With balanced weights, the unified plithogenic distance produces values $d_{\mathcal{P}} \geq 0.30$ for every Absorption case in his data, far above any reasonable threshold for "identical evaluation". The Absorption Problem is, in this sense, an artifact of dropping the attribute and contradiction terms from the metric.

Corollary 4.4 (Sufficiency of the plithogenic projection)**.** *The scalar projection π is a sufficient statistic for global calibration but not for fine-grained epistemic discrimination. The plithogenic structure $\mathcal{P}$ is sufficient for both.*

5 Comparison with Alternative Frameworks

A natural objection to the plithogenic embedding is that other multi-valued or multi-attribute frameworks could equally accommodate declared losses. We address this by setting six criteria, motivated by S4 itself, and showing that only plithogeny satisfies their conjunction.

The six criteria are:

C1 *Hyper-truth admissibility*: the framework must allow $T + I + F > 1$.

C2 *Explicit attribute spectrum*: the framework must give first-class status to a structured set of attributes V, not collapse them into a scalar.

C3 *Contradiction function*: the framework must provide a binary semantic relation c on attribute values.

C4 *Compositional closure*: the framework must admit operators (intersection, union, negation) closed under the representation.

C5 *Graded quantification*: the framework must support a graded $[0,1]^n$ membership rather than a binary classification.

C6 *Connection to dynamic epistemic state*: the framework must be amenable to time-indexed extension for reasoning chains.

Table 1 summarizes the comparison. The two-character entry "✓" denotes full satisfaction; "∼" partial; "×" failure.

Framework	C1	C2	C3	C4	C5	C6
Plithogenic neutrosophic	✓	✓	✓	✓	✓	✓
Dempster–Shafer evidence	×	×	×	✓	✓	∼
Type-2 fuzzy sets	×	×	×	✓	✓	×
Modal logic (S4, S5)	n/a	×	×	✓	×	✓
Rough sets	×	∼	×	✓	×	×
Case-based reasoning	×	∼	×	×	×	∼
Atanassov intuitionistic	×	×	×	✓	✓	×

Table 1: Conjunction of constraints. Only plithogenic neutrosophic logic satisfies C1–C6 simultaneously.

We elaborate the failures briefly.

Dempster–Shafer. Belief and plausibility are bounded by unity, so $T + I + F \leq 1$ on the standard mass assignment. The frame of discernment is unstructured and admits no contradiction function between hypotheses.

Type-2 fuzzy. The hallmark is uncertainty over membership itself, which is one-dimensional. There is no native support for $T + I + F > 1$ nor for an attribute spectrum.

Modal logic. Operators $\Box$ and $\Diamond$ classify necessity and possibility but do not graduate them. The hyper-truth question does not arise.

Rough sets. Upper and lower approximations partition the universe but assume crisp set membership. Indeterminacy is representational granularity, not an independent epistemic dimension.

Case-based reasoning. Heuristic and not axiomatized; similarity is application-specific and does not generalize.

Atanassov intuitionistic fuzzy sets. The constraint $T + F \leq 1$ explicitly excludes hyper-truth, and indeterminacy is a derived hesitancy $\pi = 1 - T - F$, not an independent dimension.

The conjunction of C1–C6 is a tight specification, and the absence of even one criterion suffices to rule out a candidate. Plithogeny satisfies all six by construction.

6 Connections with LED, N-ACH, NeutroXAI, GFD-N, and NDBG

We now show that several frameworks introduced in our prior work are either instances or extensions of the plithogenic structure of §2, and that Mason's S4 connects naturally with each.

6.1 Dynamic Epistemic Logic (LED)

LED leyvaLED formalizes the epistemic state of an LLM along a reasoning chain as a sequence $((T, I, F)_t)_{t=1}^{T}$ together with refinement, conflict, and resolution operators ρ, κ, σ acting on adjacent states. Two metrics are defined: the Neutrosophic Chain Coherence (NCC), which measures across-step compatibility, and the Resolution Index (IR), which quantifies reduction of I along the chain.

In the present setting, declared losses become typed transitions in the chain. A loss ℓ_j with high $\mathsf{severity}_j$ *increases* I_{t+1}; a resolved loss decreases it. The plithogenic attribute spectrum V_t

at step t records which losses are still open. NCC, in plithogenic notation, becomes

$$\mathrm{NCC}(t, t{+}1) = 1 - d_{\mathcal{P}}(\mathcal{P}_t, \mathcal{P}_{t+1}) \tag{7}$$

under the metric of Definition 4.3. Mason's static S4 is the special case $T = 1$.

6.2 Neutrosophic Analysis of Competing Hypotheses (N-ACH)

N-ACH leyvaNACH integrates the Admiralty Code of source weighting with cell-level (T, I, F) marking for competing hypotheses. A natural reading of S4 is that each declared loss ℓ_j is itself a competing hypothesis about the source of the model's indeterminacy. The `why` field plays the role of source justification; the `severity` field combines analyst confidence and source reliability into a scalar weight that maps directly to the Admiralty rubric. Mason's models thus perform a small N-ACH per statement, even if implicitly.

6.3 NeutroXAI

NeutroXAI leyvaNeutroXAI extends explainable AI under neutrosophic logic with component-wise uncertainty attribution. In that framework, an explanation is not a single saliency map but a plithogenic decomposition of the model's epistemic state into attributes, each with its own (T_v, I_v, F_v). S4 instantiates NeutroXAI at the explanation-output layer with two restrictions: the attributes are model-declared rather than externally fixed, and severity is scalar rather than neutrosophic. The S4-N protocol of Corollary 3.3 removes both restrictions and recovers NeutroXAI's full expressiveness.

6.4 Generative Functional Decomposition (GFD-N)

GFD-N leyvaGFDN decomposes generative behavior into attribute-indexed sub-functions. The plithogenic structure (P, v, V, d, c) is its formal backbone: V enumerates the sub-functions, d records their per-component neutrosophic performance, and c encodes inter-function semantic dependence. Declared losses populate the attribute spectrum at evaluation time and turn GFD-N into an empirically grounded model of generative uncertainty.

6.5 Neutrosophic Dynamic Bayesian Graph (NDBG)

NDBG leyvaNDBG represents temporal neutrosophic belief propagation over a graph of nodes and edges. Plithogenic attributes become nodes; the contradiction function c becomes a weight on edges, encoding semantic dissonance; per-attribute severities or (T, I, F) triples become time-indexed node states. Mason's cross-stimulus loss matrix (his Figure 3) is, structurally, an NDBG adjacency under the lexical c-estimator.

Hierarchy of epistemic fidelity. The picture that emerges is a strict inclusion ladder:

$$\begin{aligned}
&\text{Probabilistic } (T + I + F = 1) \\
&\quad \subset \text{Scalar Neutrosophic } (T, I, F) \\
&\quad \subset \text{S4 Tensor (Mason 2026): } (T, I, F) + \text{losses}_{\text{scalar severity}} \\
&\quad \subset \text{Plithogenic Neutrosophic: } (P, v, V, d, c) \\
&\quad \subset \text{LED Dynamic: } \big(\mathcal{P}_t\big)_t.
\end{aligned} \tag{8}$$

Each level is a strict refinement of the previous and each level preserves the discriminations of the level below.

7 Beyond the Unit Cube: Overset, Underset, and Offset Regimes

The plithogenic embedding developed so far has assumed $T, I, F \in [0, 1]$. smarandache2016offset generalised neutrosophic structures to three regimes that extend the unit cube and are particularly relevant to the empirical audit of LLM outputs.

Definition 7.1 (Neutrosophic overset / underset / offset, smarandache2016offset)**.** *Let $\Psi \leq 0$ and $\Omega \geq 1$. A neutrosophic evaluation $n(s) = (T, I, F)$ is said to lie in the* overset *regime if at least one component exceeds 1;* underset *regime if at least one component falls below 0;* offset *regime if both occur simultaneously across the three components. The bounded membership function $d : P \times V \to [0, 1]^3$ of Definition 2.1 is replaced by an extended membership function $d : P \times V \to [\Psi, \Omega]^3$.*

The motivating real-world examples in Smarandache's exposition are striking in their simplicity: an employee who works *overtime* deserves a degree of membership to the company strictly greater than 1, with respect to a regular full-time employee whose membership is normalised to 1; conversely, an employee whose net contribution is negative (more damage than benefit) deserves a degree of membership strictly less than 0. The standard fuzzy and neutrosophic restriction to $[0, 1]$ cannot represent either situation.

We argue that these extensions are not curiosities for LLM epistemic audit, but a natural fit for two empirically observable regimes that mason2026's S4 cannot represent within $[0, 1]^3$.

Overtime-truth (overset). A reasoning agent endowed with extended test-time compute, retrieval augmentation, or task-specific fine-tuning may declare evidence so strongly that its truth or falsity component exceeds the nominal upper bound of 1, with respect to a baseline-model evaluation calibrated to 1. The analogue of the overtime worker is a model that overcommits to a position with stronger-than-baseline support: formally, $T_v(p) > 1$ in the membership tensor of Theorem 3.1.

Adversarial truth (underset). A model under jailbreak, prompt injection, or aligned with adversarial intent produces output whose contribution to the truth-value of a statement is *negative* with respect to a baseline assessor. The standard neutrosophic framework cannot distinguish "the model is uninformed" (low T in $[0, 1]$) from "the model is actively misleading" ($T < 0$). The analogue of the damaging worker is a model whose declared truth subtracts trust rather than adding it.

Offset regime. The combined regime arises naturally when a model is partially overcommitted on one attribute and partially adversarial on another — for example, hyper-confident on a spurious sub-claim while subtly misleading on the main claim. Formally, there exist $v_i, v_j \in V$ such that $T_{v_i} > 1$ and $T_{v_j} < 0$ in the same plithogenic structure $\mathcal{P}_s$.

Proposition 7.2 (Plithogenic embedding extends to the offset regime)**.** *The plithogenic embedding of Theorem 3.1 extends without modification to the overset/underset/offset regime: the per-attribute membership tensor takes values in $[\Psi, \Omega]^3$ in place of $[0, 1]^3$; the contradiction function $c : V \times V \to [0, 1]$ is unchanged because it is a normalised semantic distance, not a membership degree; and the unified plithogenic distance $d_{\mathcal{P}}$ of Definition 4.3 is unchanged in form. The hyper-truth condition $T + I + F > 1$ generalises to the signed indicator $\sigma(s) = (T + I + F) - 1 \in \mathbb{R}$, with $\sigma > 0$ indicating overtime/conflict regimes and $\sigma < 0$ indicating adversarial/sub-baseline regimes.*

Proof. Substitute $[0, 1]^3$ by $[\Psi, \Omega]^3$ in the definitions of d and $\mathcal{P}_s$. The plithogenic operators remain closed under the extended cube because the underlying t-norms and t-conorms admit extensions to general bounded intervals (see smarandache2016offset, §3). The contradiction

function c takes values in $[0,1]$ by construction of the lexical or semantic estimator and is not affected by the change of range of d. The metric $d_{\mathcal{P}}$ depends on d only through the scalar Manhattan term d_{esc} (which extends naturally to a signed Manhattan on $\mathbb{R}^3$) and the cardinality-based Jaccard and contradiction terms (which are unchanged). The signed indicator $\sigma(s)$ is positive in the overset regime by hypothesis and negative in the underset regime, recovering the unsigned condition $\sigma > 0$ as the special case of standard hyper-truth. □

Remark 7.3 (Empirical operationalisation: the S4-O protocol)**.** *The empirical detection of overset and underset regimes in LLM output is non-trivial because models trained on probabilistic targets tend to clamp outputs to* $[0,1]$ *regardless of the prompt specification. We propose a family of four prompt strategies, collectively denoted S4-O, that anchor the extended interval explicitly through baseline references and worker-style analogies. The four strategies — direct overset elicitation, direct underset elicitation, peer-evaluation offset, and full plithogenic per-attribute offset — are reproduced verbatim in Appendix B. The peer-evaluation strategy (Appendix B.3) is of special methodological interest because it bypasses the self-report ceiling: a model rarely self-attributes adversarial status, but it can recognise adversarial output produced by a different model.*

8 Discussion

8.1 Routing on attributes generalizes routing on losses

mason2026 closes with a recommendation that AI evaluation systems route claims to evaluators *by losses rather than by capabilities*, since losses carry richer information about what an evaluator cannot do and why. Our framework gives this recommendation a formal restatement: route by the plithogenic attribute spectrum V. The advantage is that the routing decision is no longer made on a flat list of declared losses but on a structured object that also encodes the contradiction function c between losses. Routing becomes a plithogenic similarity search.

8.2 The internal distinction is the plithogenic structure

Mason writes that "the model has the distinction; the scalar cannot express it". We propose a formal name for what the model has: a single-valued plithogenic neutrosophic structure $\mathcal{P}_s$ such that $\pi(\mathcal{P}_s) = (T, I, F)$. The S4 output is a partial recovery of $\mathcal{P}_s$ at the attribute-and-severity level. The S4-N output of Corollary 3.3 would recover $\mathcal{P}_s$ in full.

8.3 Implications for attestation systems

The architectural implication for systems that record LLM evaluations for downstream audit is that the storage primitive should be the plithogenic structure, not the scalar triple plus loss list as a flat record. The contradiction function c is essential for post-hoc analysis; storing only the lexical content of `what` and `why` fields is insufficient when later evaluators may disagree on synonymy.

9 Limitations

We make four limitations explicit.

First, this paper is theoretical. We perform no new experiments; the empirical foundation is the data Mason has released. A follow-up study employing the S4-N protocol on Mason's stimuli set would test whether per-attribute neutrosophic severities reveal further distinctions that scalar severity collapses.

Second, the contradiction function c is operationalized in two ways (lexical and semantic) but not canonized. Different choices yield different plithogenic distances; we recommend reporting both c^{lex} and c^{sem} in future work, with sensitivity analysis on the weighting α, β, γ in (6).

Third, the embedding of S4 into plithogeny is at the level of single statements. Cross-statement effects, of the kind that NDBG models naturally, await empirical validation on multi-turn dialogue or chain-of-thought reasoning data.

Fourth, the comparison in §5 is structural. We do not exclude that a careful extension of, say, Dempster–Shafer with a non-standard frame structure could mimic part of the plithogenic expressiveness. Our argument is that the natural, literature-standard versions of the alternative frameworks fail the conjunction of C1–C6, and that plithogeny satisfies it without modification.

10 Conclusion

The empirical advance of mason2026 stands. Cross-vendor hyper-truth confirms that scalar neutrosophic evaluation captures a real and general property of current LLMs, and the Absorption Problem identifies a real and general limitation. We have argued that the tensor extension Mason proposes is, in formal terms, a plithogenic neutrosophic structure of single value, and that the Absorption Problem dissolves once the projection from plithogeny to the scalar triple is recognized as non-injective.

The hierarchy of epistemic fidelity is therefore: scalar (T, I, F) is necessary; the S4 tensor is an intermediate step that recovers attribute information at scalar severity; the full plithogenic structure (P, v, V, d, c) is complete at evaluation time; and LED is complete in time. Each level is required for some applications and insufficient for others, and the choice of level is itself an architectural decision.

We close with an invitation. The S4-N protocol — in which severity is reported as a neutrosophic triple per attribute, and the contradiction function c is exposed as part of the model's output — is, in our view, the natural next step in the empirical line of mason2026. The repository `github.com/fsgeek/neutrosophic-llm-logic` is a strong base on which to build it. We welcome collaboration toward this end.

Acknowledgements

The authors thank Tony Mason for releasing his data, code, and prompts openly, which made this theoretical analysis possible and grounded it in a concrete empirical record.

A Appendix A: Mason's S4 Prompt

For completeness we reproduce the S4 prompt of mason2026, which appends a list of declared losses with scalar severity to the standard (T, I, F) output. The full text is available at `github.com/fsgeek/neutrosophic-llm-logic` and is described informally in mason2026, §4.

B Appendix B: The S4-O Family of Prompts

This appendix provides the four prompt templates of the S4-O protocol introduced in Remark 7.3. The four templates share the property that the extended interval $[\Psi, \Omega]$ is anchored explicitly through reference points and Smarandache's worker analogies, in order to override the default tendency of LLMs trained on probabilistic targets to clamp outputs to $[0, 1]$.

B.1 S4-O.A — Overset elicitation with baseline anchor

> ***System.*** *You are an expert evaluator in extended Neutrosophic Logic, working in the OVERSET regime. Truth (T), Indeterminacy (I), and Falsity (F) are three independent*

dimensions, each on the EXTENDED interval $[0.0, \Omega]$ *with* $\Omega = 2.0$*. The standard reference point is* 1.0*:* $T = 1.0$ *corresponds to the truth-support that a well-calibrated baseline model would provide;* $T > 1.0$ *corresponds to stronger truth-support than the baseline (e.g., the answer is anchored in extended chain-of-thought, retrieval-augmented evidence, or domain fine-tuning);* $T < 1.0$ *corresponds to weaker truth-support than the baseline. Analogy from smarandache2016offset: an employee working overtime deserves a degree of membership* > 1 *with respect to a regular full-time employee at* 1.0*. Apply the same logic here. The three dimensions are independent and not constrained to sum to any value. You MUST stay within* $[0.0, 2.0]$*. Respond with ONLY a JSON object.*

User. *Evaluate this statement on three independent dimensions on* $[0.0, 2.0]$*: Statement: "{statement}". T (Truth): degree of truth, with* 1.0 *as baseline; I (Indeterminacy): degree of inherent uncertainty, with* 1.0 *as baseline indeterminacy; F (Falsity): degree of falsity, with* 1.0 *as baseline. Respond with ONLY:* `{"T": <v>, "I": <v>, "F": <v>}`*.*

B.2 S4-O.B — Underset elicitation with adversarial anchor

System. *You are an expert evaluator in extended Neutrosophic Logic, working in the UNDERSET regime. Truth (T), Indeterminacy (I), and Falsity (F) are three independent dimensions, each on the EXTENDED interval* $[\Psi, 1.0]$ *with* $\Psi = -1.0$*:* $T = 1.0$ *corresponds to fully truth-supported;* $T = 0.0$ *corresponds to no truth-support;* $T < 0.0$ *corresponds to actively MISLEADING — the output subtracts trust rather than adding it (e.g., known jailbreak output, prompt-injection victim, deliberately deceptive content). Analogy from smarandache2016offset: an employee who produces more damage than benefit deserves a degree of membership* < 0 *with respect to a beneficial employee at a positive degree. Negative values are reserved for adversarial regimes and should be rare. Respond with ONLY a JSON object.*

User. *You are auditing the following statement (or model output) for trust-impact. Statement: "{statement}". T (Truth):* $[-1.0, 1.0]$ *(negative = the output actively misleads); I (Indeterminacy):* $[-1.0, 1.0]$ *(negative = the output erases distinctions that should exist); F (Falsity):* $[-1.0, 1.0]$ *(negative = anti-falsifies, e.g., implies things are false that are not). Respond with ONLY:* `{"T": <v>, "I": <v>, "F": <v>}`*.*

B.3 S4-O.C — Peer-evaluation offset (recommended)

System. *You are an expert evaluator in extended Neutrosophic Logic, working in the OFFSET regime. You receive (a) a STATEMENT and (b) the OUTPUT produced by another LLM about that statement. Your task is to evaluate the trust-impact of the OTHER MODEL's output on three independent dimensions, each on* $[\Psi, \Omega]$ *with* $\Psi = -1.0$ *and* $\Omega = 2.0$*. Anchor at* 1.0*: the other model produced a baseline-correct, well-calibrated assessment. Anchor at* 0.0*: the other model contributed neither support nor anti-support.* $T, I, F > 1.0$*: OVERTIME regime — the other model gave stronger evidence than a baseline (verified citation, extended reasoning chain).* $T, I, F < 0.0$*: ADVERSARIAL regime — the other model actively misled, contradicted established facts, or was jailbroken. Respond with ONLY a JSON object.*

User. *Statement: "{statement}". Output produced by another model: "{other_model_output}". Evaluate the OTHER MODEL's output on the OFFSET interval* $[-1.0, 2.0]$*. If the other output exceeds baseline calibration on a dimension, report* > 1.0*. If it actively misleads on a dimension, report* < 0.0*. Otherwise stay in* $[0.0, 1.0]$*. Respond with ONLY:* `{"T": <v>, "I": <v>, "F": <v>}`*.*

The peer-evaluation strategy bypasses the self-report ceiling that frustrates the direct elicitation of underset values: a model rarely self-attributes adversarial status, but it can recognise adversarial output produced by a different model. Empirical validation of B.3 on Mason's stimuli set is the natural follow-up empirical study.

B.4 S4-O.D — Plithogenic per-attribute offset

System. *You are an expert evaluator in plithogenic neutrosophic logic with overset/underset/offset extensions. For each statement, you produce: (1) a scalar* $(T, I, F) \in [-1, 2]^3$ *as global trust-impact; (2) a list of attribute-specific evaluations, where each attribute is declared by you*

(free-form "what" field) and carries its own per-attribute $(T_v, I_v, F_v) \in [-1, 2]^3$. *Anchors:* $+1.0$ *baseline-calibrated;* 0.0 *no contribution;* > 1 *overtime regime (stronger than baseline);* < 0 *adversarial regime (actively misleading). Respond with ONLY a JSON object.*

User. *Statement: "{statement}". Provide* `{"scalar": {"T":..., "I":..., "F":...}, "attributes": [{"what":..., "why":..., "T_v":..., "I_v":..., "F_v":...}, ...]}`. *Use overset values (*> 1*) where you have stronger-than-baseline evidence; use underset values (*< 0*) where the corresponding aspect of the statement is actively misleading.*

This last template closes the loop with the plithogenic structure (P, v, V, d, c): each declared attribute is a member of V with its own neutrosophic membership in the offset cube $[-1, 2]^3$. The contradiction function c between attributes can then be estimated by the lexical or semantic estimator of Definition 4.3.

Practical considerations. Three risks should be reported in any empirical use of S4-O. *First*, models may clamp outputs to $[0, 1]$ regardless of the prompt instructions; the extent of clamping should be measured as a side outcome of the experiment. *Second*, alignment-tuned models may refuse to report $T < 0$ on safety grounds; the audit framing of B.3 partially mitigates this. *Third*, overset inflation (artificially large T values without commensurate justification) should be controlled by requiring the "why" field in B.4 and validating consistency across repetitions.

References

[Aamodt and Plaza, 1994] Aamodt, A., and Plaza, E. (1994). Case-based reasoning: foundational issues, methodological variations, and system approaches. *AI Communications*, 7(1):39–59.

[Atanassov, 1986] Atanassov, K. (1986). Intuitionistic fuzzy sets. *Fuzzy Sets and Systems*, 20(1):87–96.

[Baltag, Moss, and Solecki, 1998] Baltag, A., Moss, L. S., and Solecki, S. (1998). The logic of public announcements, common knowledge, and private suspicions. *Proceedings of TARK 1998*, 43–56.

[van Benthem, 2011] van Benthem, J. (2011). *Logical Dynamics of Information and Interaction*. Cambridge University Press.

[Dempster, 1968] Dempster, A. P. (1968). A generalization of Bayesian inference. *Journal of the Royal Statistical Society B*, 30(2):205–247.

[Hintikka, 1962] Hintikka, J. (1962). *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Cornell University Press.

[Leyva-Vázquez, 2024a] Leyva-Vázquez, M. (2024). Dynamic epistemic neutrosophic logic for LLM reasoning chains: NCC and IR metrics. *Neutrosophic Sets and Systems*, in press.

[Leyva-Vázquez, 2024b] Leyva-Vázquez, M. (2024). Neutrosophic analysis of competing hypotheses with Admiralty weighting. *Neutrosophic Computing and Machine Learning*, in press.

[Leyva-Vázquez, 2024c] Leyva-Vázquez, M. (2024). NeutroXAI: explainability under neutrosophic logic with component-wise attribution. *Neutrosophic Sets and Systems*, in press.

[Leyva-Vázquez, 2024d] Leyva-Vázquez, M. (2024). Generative functional decomposition under neutrosophic uncertainty (GFD-N). *Neutrosophic Sets and Systems*, in press.

[Leyva-Vázquez, 2024e] Leyva-Vázquez, M. (2024). Neutrosophic dynamic Bayesian graphs for temporal belief propagation. *Neutrosophic Sets and Systems*, in press.

[Leyva-Vázquez and Smarandache, 2025] Leyva-Vázquez, M., and Smarandache, F. (2025). Breaking the chains of probability: neutrosophic logic as a new framework for epistemic uncertainty in large language models. Preprint.

[Mason, 2026] Mason, T. (2026). From scalars to tensors: declared losses recover epistemic distinctions that neutrosophic scalars cannot express. *arXiv:2604.09602v1 [cs.AI]*.

[Mendel and John, 2002] Mendel, J. M., and John, R. I. B. (2002). Type-2 fuzzy sets made simple. *IEEE Transactions on Fuzzy Systems*, 10(2):117–127.

[Pawlak, 1982] Pawlak, Z. (1982). Rough sets. *International Journal of Computer and Information Sciences*, 11(5):341–356.

[Rivieccio, 2008] Rivieccio, U. (2008). Neutrosophic logics: prospects and problems. *Fuzzy Sets and Systems*, 159(14):1860–1868.

[Shafer, 1976] Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press.

[Shannon, 1948] Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

[Smarandache, 2005] Smarandache, F. (2005). *A Unifying Field in Logics: Neutrosophic Logic. Neutrosophy, Neutrosophic Set, Neutrosophic Probability*. American Research Press, Rehoboth.

[Smarandache, 2010] Smarandache, F. (2010). Neutrosophic set, a generalization of the intuitionistic fuzzy set. *Journal of Defense Resources Management*, 1(1):107–116.

[Smarandache, 2016] Smarandache, F. (2016). *Neutrosophic Overset, Neutrosophic Underset, and Neutrosophic Offset. Similarly for Neutrosophic Over-/Under-/Off- Logic, Probability, and Statistics*. Pons Editions, Brussels. Available at https://fs.unm.edu/NeutrosophicOversetUndersetOffset.pdf.

[Smarandache, 2018] Smarandache, F. (2018). *Plithogenic Set: An Extension of Crisp, Fuzzy, Intuitionistic Fuzzy, and Neutrosophic Sets — Revisited. Neutrosophic Sets and Systems*, 21:153–166.

[Smarandache, 2019] Smarandache, F. (2019). Plithogenic logic and operators. *International Journal of Neutrosophic Science*, 1(2):69–80.

[Smarandache, 2023] Smarandache, F. (2023). Foundations of HyperUncertainty in plithogeny: a unified framework. *Neutrosophic Sets and Systems*, 56:1–18.

[Yadkori et al., 2024] Abbasi Yadkori, Y., Kuzborskij, I., Stutz, D., György, A., Fisch, A., Doucet, A., et al. (2024). Mitigating LLM hallucinations via conformal abstention. *arXiv:2405.01563*.

[Tayebati et al., 2025] Tayebati, S., Kumar, D., Darabi, N., Jayasuriya, D., Krishnan, R., and Trivedi, A. R. (2025). Learning conformal abstention policies for adaptive risk management in large language and vision-language models. *arXiv:2502.06884*.

[Machcha et al., 2026] Machcha, S., Yerra, S., Gupta, S., Sahoo, A., Sultana, S., Yu, H., and Yao, Z. (2026). Knowing when to abstain: medical LLMs under clinical uncertainty. *arXiv:2601.12471*.

[Li et al., 2024] Li, S., Yang, C., Wu, T., Shi, C., Zhang, Y., Zhu, X., et al. (2024). A survey on the honesty of large language models. *arXiv:2409.18786*.

[Yan et al., 2026] Yan, S.-Q., Li, Y., Liu, Q., and Ling, Z.-H. (2026). Learn to be honest: mitigate LLMs' overconfidence for improving hallucination detection with self-hesitation activation. Preprint.

[Zadeh, 1975] Zadeh, L. A. (1975). The concept of a linguistic variable and its application to approximate reasoning. *Information Sciences*, 8(3):199–249.